PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04J 14/02

A1

(11) International Publication Number: WO 00/04667

(43) International Publication Date: 27 January 2000 (27.01.00)

(21) International Application Number: PCT/US99/14979

(22) International Filing Date: 1 July 1999 (01.07.99)

(30) Priority Data:
09/118,437 17 July 1998 (17.07.98) US

(71) Applicant: TELCORDIA TECHNOLOGIES, INC. [US/US]; 445 South Street, Morristown, NJ 07960–6438 (US).

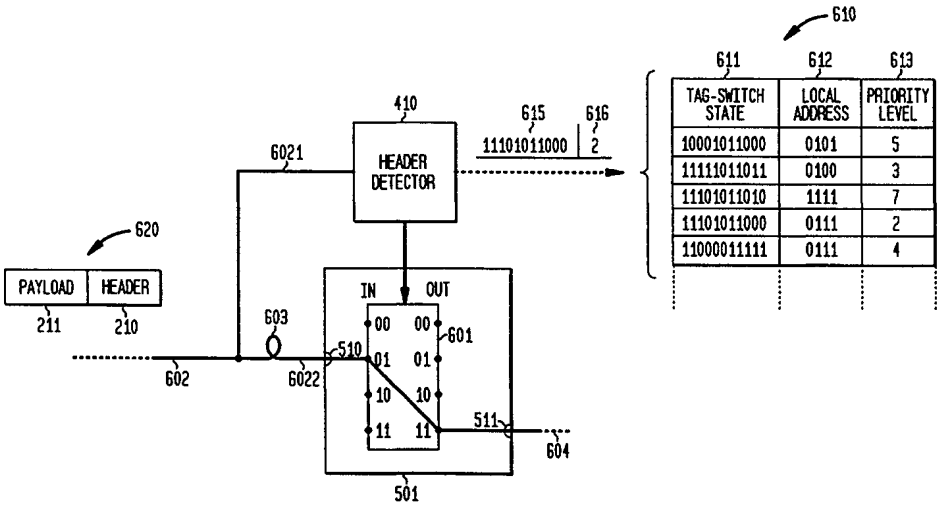
(72) Inventors: CHANG, Gee–Kung; 7 East Lawn Drive, Holmdel, NJ 07733 (US). YOO, Sung–Joo; 635 Cleveland Street, Davis, CA 95616 (US).

(74) Agents: GIORDANO, Joseph et al.; International Coordinator, Rm. 1G112R, 445 South Street, Morristown, NJ 07960–6438 (US).

(81) Designated States: AU, CA, CN, ID, IN, JP, KR, MX, SG, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: HIGH–THROUGHPUT, LOW–LATENCY NEXT GENERATION INTERNET NETWORKS USING OPTICAL TAG SWITCHING



(57) Abstract

An optical signaling header (210) technique applicable to optical networks wherein packet (620) routing information is embedded in the same channel or wavelength as the data payload (211) so that both the header (210) and data (211) payload propagate through network elements with the same path and the associated delays. The header (210) information has sufficiently different characteristics from the data payload (211) so that the signaling header can be detected without being affected by the data payload, and that the signaling header can also be removed without affecting the data payload. The signal routing technique can overlaid onto the conventional network elements in a modular manner using two types of applique modules. The first type effects header encoding and decoding at the entry and exit points of the data payload into and out of the network; the second type effects header detection at each of the network elements.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

HIGH-THROUGHPUT, LOW-LATENCY NEXT GENERATION INTERNET NETWORKS USING OPTICAL TAG SWITCHING

BACKGROUND OF THE DISCLOSURE

1. Field of the Invention

This invention relates to optical communication systems and, more particularly, to an optical system which accommodates network traffic with high throughput and low latency.

2. Description of the Background Art

Recent research advances in optical Wavelength Division Multiplexing (WDM) technology have fostered the development of networks that are orders of magnitude higher in transmission bandwidth than existing commercial networks. While such an increase in throughput is impressive on its own, a corresponding decrease in network latency must also be achieved in order to realize the Next Generation Internet (NGI) vision of providing the next generation of ultra high speed networks that can meet the requirements for supporting new applications, including national initiatives. Towards this end, current research efforts have focused on developing an ultra-low latency Internet Protocol (IP) over WDM optical packet switching technology that promises to deliver the two-fold goal of both high throughput with low latency. Such efforts, while promising, have yet to fully realize this two-fold goal.

There are a number of challenging requirements in realizing such IP/WDM networks. First, the NGI network must inter-operate with the existing Internet and avoid protocol conflicts. Second, the NGI network must provide not only ultra low-latency, but must take advantage of both packet-switched (that is, bursty) IP traffic and circuit-switched WDM networks. Third, it is advantageous if the NGI network requires

no synchronization between signaling and data payload. Finally, a desired objective is that the NGI network accommodates data traffic of various protocols and formats so that it is possible to transmit and receive IP as well as non-IP signals without the need for complicated synchronization or format conversion.

Comparison with other work

The Multi-Wavelength-Optical Network (MONET) system, as reported in the article "MONET: Multi-Wavelength Optical Networking" by R. E. Wagner, et al. and published in the Journal of Lightwave Technology, Vol. 14, No. 6, June 1996, demonstrated a number of key milestones in optical network including transparent transmission of multi-wavelength through more than 12 reconfigurable network elements spread over the national scale fiber distance. The network, however, is circuit-switched and suffers inefficiency in accommodating bursty traffic. The typical connection setup time from request to switching is a few seconds, limited by capabilities of both Network Control & Management (NC&M) and hardware. Recent efforts within the MONET program to improve on the efficiency concentrated on the "Just-in-Time signaling" scheme. This method utilizes embedded 1510 nm NC&M signaling which precedes the data payload by an estimated delay time. This estimation must be accurately made for each network configuration for every wavelength in order to synchronize the signaling header and switching of the payload.

In accordance with the present invention, the optical packet header is carried over the same wavelength as the packet payload data. This approach eliminates the issue of header and payload synchronization. Furthermore, with a suitable use of optical delay at each intermediate optical switch, it eliminates the need to estimate the initial burst delay by incorporating the optical delay directly at the switches. This makes

a striking difference with Just-In-Time signaling in which the delay at each switch along the path needs to be known ahead of time and must be entered in the calculation for the total delay. Lastly, there is little time wasted in requesting a connection time and actually achieving a connection. In comparison to a few second delays seen in MONET, the present inventive subject matter reduces the delay to minimal, only limited by the actual hardware switching delays at each switch. The current switching technology realizes delays of only several microseconds, and shorter delays will be possible in the future. Such a short delay can be incorporated by using an optical fiber delay line at each network element utilizing switches. The present inventive subject matter achieves the lowest possible latency down to the fundamental limit of the hardware, and no lower latency can be achieved by any other technique.

The Optical Networks Technology Consortium (ONTC) results were reported in the article "Multiwavelength Reconfigurable WDM/ATM/SONET Network Testbed" by Chang et al. and published in the Journal of Lightwave Technology, Vol. 14, No. 6, June 1996. Both Phase I (155 Mb/s, 4-wavelength) and Phase II (2.5 Gb/s, 8-wavelength) of the ONTC program were configured on a Multihop ATM-based network. While such an ATM based architecture added a large overhead and excluded the possibility of a single-hop network, the packet/header signaling was made possible by utilizing the isochronous ATM cell itself. This communication of NC&M information is made through the same optical wavelength, potentially offering similar benefits as with the technique of the present invention. However, the inventive technique offers a number of significant advantages over the ATM-based signaling. First, the inventive technique offers a single hop connection without the need to convert to electrical signals and buffer the packets. Second, it offers far more efficient utilization of the bandwidth by

eliminating excessive overheads. Third, it allows strictly transparent and ultra-low latency connections.

The ARPA sponsored All-Optical-Network (AON) Consortium results were reported in an article entitled "A Wideband All-Optical WDM Network" , by I. P. Kaminow et al. and published in the IEEE Journal on Selected Areas of Communication, Vol. 14, No. 5, June, 1996. There were actually two parts of the AON program: WDM as reported in the aforementioned article, and TDM reported in a companion paper in the same issue. First the WDM part of the AON program is first discussed, followed by the TDM part.

The AON architecture is a three-level hierarchy of subnetworks, and resembles that of LANs, MANs, and WANs seen in computer networks. The AON provides three basic services between Optical Terminals (OTs): A, B, and C services. A is a transparent circuit-switched service, B is a transparent time-scheduled TDM/WDM service, and C is a non-transparent datagram service used for signaling. The B service uses a structure where a 250 msec frame is used with 128 slots per frame. Within a slot or group of slots, a user is free to choose the modulation rate and format. The B-service implemented on the AON architecture is closest to the IP over WDM which is the subject matter of the present invention. However, the separation of NC&M signaling in the C-service with the payload in the B-service requires careful synchronization between the signaling header and the payload. This requirement becomes far more stringent as the 250 microsecond frame is used with 128 slots per frame with arbitrary bit rates. Not only the synchronization has to occur at the bit level, but this synchronization has to be achieve across the entire network. The scalability and interoperability are extremely difficult since these do not go in steps with the network synchronization requirement.

The present inventive subject matter requires no synchronization, inter-operates with existing IP and non-IP traffic, and offers scalability.

TDM efforts are aimed at 100 Gb/s bit rates. In principle, such ultrafast TDM networks have the potential to provide truly flexible bandwidth on demand at burst rates of 100 Gb/s. However, there are significant technological challenges behind such high bit rate systems mainly related to nonlinearities, dispersion, and polarization degradations in the fiber. While the soliton technologies can alleviate some of the difficulties, it still requires extremely accurate synchronization of the network -- down to a few picoseconds. In addition, the header and the payload must have the identical bit rates, and as a consequence, bit-rate transparent services are difficult to provide. The subject matter in accordance with the present invention requires no synchronization, relies on no 100 Gb/s technologies, and offers transparent services.

The Cisco Corporation recently announced a product based on Tag-Switching and the general description of Cisco's Tag-Switching is available at the world-wide-web site, (http://www.cisco.com/warp/public/732/tag/). Cisco's (electronic) Tag Switching assigns a label or "tag" to packets traversing a network of routers and switches. In a conventional router network, each packet must be processed by each router to determine the next hop of the packet toward its final destination. In an (electronic) Tag Switching network, tags are assigned to destination networks or hosts. Packets then are switched through the network with each node simply swaps tags rather than processing each packet. An (electronic) Tag Switching network will consist of a core of (electronic) tag switches (either conventional routers or switches), which connect to (electronic) tag edge routers on the network's periphery. (Electronic) Tag edge routers and tag switches use standard routing protocols to identify routes through the network. These systems

then use the tables generated by the routing protocols to assign and distribute tag information via a Tag Distribution Protocol. Tag switches and tag edge routers receive the Tag Distribution Protocol information and build a forwarding database. The database maps particular destinations to the tags associated with those destinations and the ports through which they are reachable.

When a tag edge router receives a packet for forwarding across the tag network, it analyzes the network-layer header and performs applicable network layer services. It then selects a route for the packet from its routing tables, applies a tag and forwards the packet to the next-hop tag switch.

The tag switch receives the tagged packet and switches the packet based solely on the tag, without re-analyzing the network-layer header. The packet reaches the tag edge router at the egress point of the network, where the tag is stripped off and the packet delivered. After Cisco made its announcement about (Electronic) Tag Switching, the IETF (Internet Engineering Task Force) has recommended a MPLS (Multi-protocol Label Switching) to implement standardized, vendor-neutral (electronic) tag-switching function in routers and switches, including ATM switches.

A number of features in the Cisco's (electronic) Tag Switching is similar to the Optical Tag Switching which is the subject matter of the present invention, with the features aimed at the similar goals of simplifying the processing required for packet routing. The key differences are as follows. First, the optical tag switching is purely optical in the sense that both tag and data payload are in an optical form. While each plug-and-play module (a component of the present inventive system) senses the optical tag, the actual packet does not undergo optical-to-electrical conversion until it comes out of the network The Cisco's (electronic) Tag Switching will be all electrical, and applies

electronic detection, processing, and retransmission to each packet at each router. Secondly, the Optical Tag Switching of the present invention achieves lowest possible latency and does not rely on utilizing buffers. Electronic tag switching will have far greater latency due to electronic processing and electronic buffering. Thirdly, the Optical Tag Switching of the present invention utilizes path deflection and/or wavelength conversion to resolve blocking due to contention of the packets, whereas the Electronic Tag Switching will only utilize electronic buffering as a means to achieve contention resolution at the cost of increased latency, and the performance is strongly dependent on packet size. The present invention covers packets of any length. Lastly, the Optical Tag Switching of the present invention achieves a strictly transparent network in which data of any format and protocol can be routed so long as it has a proper optical tag. Hence the data can be digital of any bit rate, analog, or FSK (frequency-shifted-keying) format. The Electronic Tag Switching requires that data payload to have the given digital bit rate identical to the electronic tag since the routers must buffer them electronically.

Another representative technology that serves as background to the present invention is the so-called Session Deflection Virtual Circuit Protocol (SDVC), which is based on deflection routing method. The paper entitled "The Manhattan Street Network", by N. F. Maxemchuk" as published in the Proceedings on IEEE Globecom '85, pp 255-261, December 1985, discusses that when two packets attempt to go to the same destination, one packet can be randomly chosen for the preferred output link and the other packet is "deflected" to the non-preferred link. This means that packets will occasionally take paths that are not shortest paths. The deflection method utilized by the present invention does not 'randomly' select the packet to go to the most preferred path; rather, it attempts to look into the priorities of the packets, and send the higher priority

packet to be routed to the preferred path. The packets will be deflected if they have lower priorities; however, both 'path deflection' and 'wavelength deflection' are utilized. The path deflection is similar to conventional SDVC in that the optical packet will be simply routed to the path of the next preference at the same wavelength. The wavelength deflection allows the optical packet to be routed to the most preferred path but at a different wavelength. This wavelength deflection is achieved by wavelength conversion at the network elements. Partially limited wavelength conversion is utilized, meaning not all wavelengths will be available as destination wavelengths for a given originating wavelength. The wavelength deflection allows resolution of blocking due to wavelength contentions without increasing the path delay. The combination of path and wavelength deflections offers sufficiently large additional connectivities for resolving packet contentions; however, the degree of partial wavelength conversion can be increased when the blocking rate starts to rise. Such scalability and flexibility of the network is not addressed by conventional SDVC.

SUMMARY OF THE INVENTION

The present invention utilizes a unique optical signaling header technique applicable to optical networks. Packet routing information is embedded in the same channel or wavelength as the data payload so that both the header and data information propagate through the network with the same path and the associated delays. However, the header routing information has sufficiently different characteristics from the data payload so that the signaling header can be detected without being affected by the data payload and that the signaling header can also be stripped off without affecting the data payload. The inventive subject matter allows such a unique signal routing method to be

overlaid onto the conventional network elements in a modular manner, by adding two types of 'Plug-and-Play' modules. The inventive subject matter overcomes the shortcomings and limitation of other methods discussed in the Background section while advantageously utilizing the full capabilities of optical networking.

In accordance with the broad method aspect of the present invention, a method for propagating a data payload from an input network element to an output network element in a wavelength division multiplexing system composed of a plurality of network elements, given that the data payload has a given format and protocol, includes the following steps: (a) generating and storing a local routing table in each of the network elements, each local routing table determining a local route through the associated one of the network elements; (b) adding an optical header to the data payload prior to inputting the data payload to the input network element, the header having a format and protocol and being indicative of the local route through each of the network elements for the data payload and the header, the format and protocol of the data payload being independent of the format and protocol of the header; (c) optically determining the header at each of the network elements as the data payload and header propagate through the WDM network; (d) selecting the local route for the data payload and the header through each of the network elements as determined by looking up the header in the corresponding local routing table; and (e) routing the data payload and the header through each of the network elements in correspondence to the selected route.

In accordance with the broad system aspect of the present invention, the system is arranged in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source in the

electrical layer and destined for a destination in the electrical layer, the data payload having a given format and protocol. The system includes: (i) a first type of optical header module, coupling the source in the optical layer and the WDM network, for adding an optical header ahead of the data payload prior to inputting the data payload to the WDM network, the header being indicative of a local route through the network elements for the data payload and the header, the format and protocol of the data payload being independent of those of the header; and (ii) a second type of optical header module, appended to each of the network elements, for storing a local routing table in a corresponding one of the network elements, each local routing table determining a routing path through the corresponding one of the network elements, for optically determining the header at the corresponding one of the network elements as the data payload and header propagate over the WDM network, for selecting the local route for the data payload and the header through the corresponding one of the network elements as determined by looking up the header in the corresponding local routing table, and for routing the data payload and the header through the corresponding one of the network elements in correspondence to the selected route.

The present invention offers numerous features and benefits, including (1) extremely low latency limited only by hardware delays; (2) high throughput and bandwidth-on-demand offered by combining multi-wavelength networking and optical tag switching; (3) priority based routing which allows higher throughput for higher priority datagrams or packets; (4) scalable and modular upgrades of the network from the conventional WDM to the inventive optical tag-switched WDM; (5) effective routing of long datagrams, consecutive packets, and even non-consecutive packets; (6) cost-effective utilization of optical components such as multiplexers and fibers; (7)

interoperability in a multi-vendor environment; (8) graceful and step-by-step upgrades of network elements; (9) transparent support of data of any format and any protocol; and (10) high quality-of-service communications.

BRIEF DESCRIPTION OF THE DRAWINGS

The teachings of the present invention can be readily understood by considering the following detailed description in conjunction with the accompanying drawings, in which:

FIG. 1 is a pictorial representation of a general network illustrating the coupling between the optical and electrical layers of the network as effected by one aspect of the present invention;

FIG. 2 illustrates the optical layer of the network of FIG. 1 showing the relationship between the optical signal header and data payload, and the use of the header/payload in network setup;

FIG. 3 is a high-level block diagram of one Plug & Play module in accordance with the present invention for header encoding and header removal;

FIG. 4 is a high-level block diagram of another Plug & Play module in accordance with the present invention for routing a packet through a WDM network element;

FIG. 5 is illustrative of a WDM circuit-switched backbone network;

FIG. 6 illustrates a network element of FIG. 1 with its embedded switch and the use of local routing tables;

FIG. 7 depicts a block diagram of an illustrative embodiment of a header encoder circuit for the Plug-&-Play module of FIG. 3;

FIG. 8 depicts a block diagram of an illustrative embodiment of a header remover circuit for the Plug-&-Play module of FIG. 3;

FIG. 9 depicts a block diagram of an illustrative embodiment of a header detector circuit for the Plug-&-Play module of FIG. 4;

FIG. 10 depicts a block diagram for a more detailed embodiment of FIG. 4 wherein the tag-switch controller includes interposed demultiplexers, and header detectors and fast memory; and

FIG. 11 is a flow diagram for the processing effected by each tag-switch controller of FIG. 10.

To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures.

DETAILED DESCRIPTION

In order to gain an insight into the fundamental principles in accordance with the present invention as well as to introduce terminology useful in the sequel, an overview is first presented, followed by an elucidation of an illustration embodiment.

Overview

The present invention relates to a network for realizing low latency, high throughput, and cost-effective bandwidth-on-demand for large blocks of data for NGI applications. Cost-effective and interoperable upgrades to the network are realized by interposing portable 'Plug-and-Play' modules on the existing WDM network elements to effect so-called "WDM optical tag switching" or, synonymously, "optical tag switching". The invention impacts both the hardware and software for the NGI network from all

perspectives, including architecture, protocol, network management, network element design, and enabling technologies.

As alluded to, the methodology carried out by the network and concomitant circuitry for implementing the network are engendered by a technique called WDM optical tag-switching -- defined as the dynamic generation of a routing path for a burst duration by an in-band optical signaling header. Data packets are routed through the WDM network using an in-band WDM signaling header for each packet. At a switching node, the signaling header is processed and the header and the data payload (1) may be immediately forwarded through an already existing flow state connection, or (2) a path can be setup for a burst duration to handle the header and the data payload. WDM tag-switching enables highly efficient routing and throughput, and reduces the number of IP-level hops required by keeping the packets routing at the optical level to one hop as managed by the NC&M which creates and maintains routing information.

The depiction of FIG. 1 shows the inter-relation between optical layer 120 and electrical layer 110 of generic network 100 as provided by intermediate layer 130 coupling the optical layer and the electrical layer. Electrical layer 110 is shown, for simplicity, as being composed of two conventional IP routers 111 and 112. Optical layer 120 is shown as being composed of network elements or nodes 121-125. Intermediate layer 130 depicts conventional ATM/SONET system 131 coupling IP router 112 to network element 122. Also shown as part of layer 130 is header network 132, which in accordance with the present invention, couples IP router 111 to network element 121. FIG. 1 pictorially illustrates the location of network 132 on a national-scale, transparent WDM-based backbone network with full interoperability and reconfigurability. It is important to emphasize at this point that the elements of FIG. 1 are illustrative of one

embodiment in accordance with the present invention; thus, for example, element 111 may, in another embodiment, be an ATM router or even a switch.

Now with reference to FIG. 2, optical layer 120 of FIG. 1 is shown in more detail including the basic technique, in accordance with the present invention, for setting up a fast connection in optical network 201, composed of network elements 121-125; the setup uses optical signaling header 210 for the accompanying data payload 211. This technique combines the advantages of circuit-switched based WDM and packet-switched based IP technologies. New signaling information is added in the form of an optical signal header 210 which is carried in-band within each wavelength in the multi-wavelength transport environment. Optical signaling header 210 is a tag containing routing and control information such as the source, destination, priority, and the length of the packet, propagates through optical network 201 preceding data payload 211. Each WDM network element 121-125 senses optical signaling header 210, looks-up a connection table (discussed later), and takes necessary steps such as cross-connections, add, drop, or drop-and-continue. The connection table is constantly updated by continuous communication between NC&M 220 and WDM network elements 121-125. Data payload 211, which follows optical signaling header 210, is routed through a path in each network element (discussed later) as established by the connection. With the arrangement of FIG. 2, there is no need to manage the time delay between optical signaling header 210 and data payload 211, shown by T in FIG. 2, because each network element provides the optical delay needed for the short time required for connection set-up within each network element via delay on an interposed fiber. Moreover, the format and protocol of the data payload is independent of that of the header, that is, for a given network whereas the format and protocol of the header are pre-determined, the format

and the protocol of the data payload can be the same as or different from those of the header.

Each destination is associated with a preferred path which would minimize 'the cost' – in FIG. 2, the overall path from source 123 to destination 122 includes paths 201 and 202 in cascade, both utilizing wavelength WP. This cost is computed based on the total propagation distance, the number of hops, and the traffic load. The preferred wavelength is defaulted to the original wavelength. For example, the preferred wavelength on path 202 is WP. If this preferred path at the default wavelength is already occupied by another packet, then network element 121 quickly decides if there is an available alternate wavelength WA through the same preferred path. This alternate wavelength must be one of the choices offered by the limited wavelength conversion in network element 121. If there is no choice of wavelengths which allows transport of the packet through the most preferred path, the next preferred path is selected (path deflection). For example, in FIG. 2, paths 203 and 204 in cascade may represent the alternative path. At this point, the preferred wavelength will default back to the original wavelength WP. The identical process of looking for an alternate wavelength can proceed if this default wavelength is again already occupied. In FIG. 2, path 203 is an alternative path with the same wavelength WP, and path 204 is an alternate path using alternate wavelength WA. In an unlikely case where there is no combination of path and wavelength deflection can offer transport of the packet, network element 121 will decide to drop the packet of lower priority. In other words, the new packet transport through the preferred path at the originating wavelength takes place by dropping the other packet of the lower priority which is already occupying the preferred path.

Network elements 121-125 are augmented with two types of so-called 'Plug-and-Play' modules to efficiently handle bursty traffic by providing packet switching capabilities to conventional circuit-switched WDM network elements 121-125 whereby signaling headers are encoded onto IP packets and are removed when necessary.

The first type of 'Plug-and-Play' module, represented by electro-optical element 132 of FIG. 1, is now shown in block diagram form in FIG. 3. Whereas conceptually module 132 is a stand-alone element, in practice, module 132 is integrated with network element 121 as is shown in FIG. 3; module 132 is interposed between compliant client interface (CCI) 310 of network element 121 and IP router 111 to encode optical signaling header 210 onto the packets added into the network via header encoder 321, and to remove optical signaling header 210 from the packets dropping out of the network via header remover 322.

Generally, encoding/removing module 132 is placed where the IP traffic is interfaced into and out of the WDM network, which is between the client interface of the network element and the IP routers. The client interfaces can be either a CCI-type or a non-compliant client interfaces (NCI)-type. At these interfaces, header encoder 321 puts optical header 210 carrying the destination and other information in front of data payload 211 as the IP signal is transported into network 201. Optical header 210 is encoded in the optical domain by an optical modulator (discussed later). Signaling header remover 322 deletes header 210 from the optical signal dropped via a client interface, and provides an electrical IP packet to IP router 111.

More specifically, module 132 accepts the electrical signal from IP router 111, converts the electrical signal to a desired compliant wavelength optical signal, and places optical header 210 in front of the entire packet. Module 132 communicates with

NC&M 220 and buffers the data before optically converting the data if requested by NC&M 220. Module 132 employs an optical transmitter (discussed later) with the wavelength matched to the client interface wavelength. (As indicated later but instructive to mention here, module 132 is also compatible with NCI 404 of FIG. 4 since the wavelength adaptation occurs in the NCI; however, the bit-rate-compatibility of NCI wavelength adaption and the IP signal with optical headers must be established in advance.)

FIG. 4 depicts a second type of 'Plug-and-Play' module, optical element 410, which is associated with each WDM network element 121-125, say element 121 for discussion purposes. Module 410 is interposed between conventional network element circuit switch controller 420 and conventional switching device 430. Module 410 detects information from each signaling header 210 propagating over any fiber 401-403, as provided to module 410 by tapped fiber paths 404-406. Module 410 functions to achieve very rapid table look-up and fast signaling to switching device 430. Switch controller 420 is functionally equivalent to the conventional "craft interface" used for controlling the network elements; however, in this case, the purpose of this switch controller 420 is to accept the circuit-switched signaling from NC&M 220 and determine which control commands are to be sent to tag switch controller 410 based on the priority. Thus, tag switch controller 410 receives circuit-switched control signals from network element circuit switch controller 420, as well as information as derived from each signaling each header 210, and intelligently choose between the circuit-switched and the tag-switched control schemes. The switches (discussed later) comprising switching device 430 also achieve rapid switching. The delay imposed by fibers 415, 416, or 416, which are placed in input paths 401-403 to switching device 430, are such that the delay is larger than the

total time it takes to read signaling header 210, to complete a table look-up, and to effect switching. Approximately, a 2 km fiber provides 10 microsecond processing time. The types of WDM network elements represented by elements 121-125 and which encompass switching device 430 include: Wavelength Add-Drop Multiplexers (WADMs); Wavelength Selective Crossconnects (WSXCs); and Wavelength Interchanging Crossconnects (WIXCs) with limited wavelength conversion capabilities.

In operation, module 410 taps a small fraction of the optical signals appearing on paths 401-403 in order to detect information in each signaling header 210, and determine the appropriate commands for switching device 430 after looking up the connection table stored in module 410. The fiber delay is placed in paths 401-403 so that the packet having header 210 and payload 211 reaches switching device 430 only after the actual switching occurs. This fiber delay is specific to the delay associated with header detection, table look-up, and switching, and can typically be accomplished in about 10 microseconds with about 2 km fiber delay in fibers 415-417.

Since there is no optical-to-electrical, nor electrical-to-optical conversion of data payload 211 at network elements 121-125, the connections are completely transparent. Contrary to IP routing, where a multiplicity of bit-rates and lower-level protocols increases the number of different interfaces required and consequently the cost of the router, routing by WDM tag switching is transparent to bit-rates. By way of illustration, optical routing by network elements 121-125 is able to achieve 1.28 Tb/sec throughput (16x16 cross-connect switching device 430 with 32 wavelengths/fiber at 2.5Gb/sec per wavelength) which is much larger than any of the current gigabit routers.

Each network element 121-125 in combination with NC&M 220 effects a routing protocol which is adaptive; the routing protocol performs the following functions:

(a) measures network parameters, such as state of communication lines, estimated traffic, delays, capacity utilization, pertinent to the routing strategy; (b) forwards the measured information to NC&M 220 for routing computations; (c) computes of the routing tables at NC&M 220; (d) disseminates the routing tables to each network element 121-125 to have packet routing decisions at each network element. NC&M 220 receives the network parameter information from each network element, and updates the routing tables periodically, then (e) forwards a connection request from an IP router such as element 111 to NC&M 220, and (f) forwards routing information from the NC&M 220 to each network element 121-125 to be inputted in optical signaling header 210.

Packets are routed through network 201 using the information in signaling header 210 of each packet. When a packet arrives at a network element, signaling header 210 is read and either the packet (a) is routed to a new appropriate outbound port chosen according to the tag routing look-up table, or (b) is immediately forwarded through an already existing tag-switching originated connection within the network element. The latter case is referred to as "flow switching" and is supported as part of optical tag-switching; flow switching is used for large volume bursty mode traffic.

Tag-switched routing look-up tables are included in network elements 121-125 in order to rapidly route the optical packet through the network element whenever a flow switching state is not set-up. The connection set-up request conveyed by optical signaling header 210 is rapidly compared against the tag-switch routing look-up table within each network element. In some cases, the optimal connections for the most efficient signal routing may already be occupied. The possible connection look up table is also configured to already provide an alternate wavelength assignment or an alternate path to route the signal. Providing a limited number of (at least one) alternative

wavelength significantly reduces the blocking probability. The alternative wavelength routing also achieves the same propagation delay and number of hops as the optimal case, and eliminates the difficulties in sequencing multiple packets. The alternate path routing can potentially increase the delay and the number of hops, and the signal-to noise-ratio of the packets are optically monitored to eliminate any possibility of packets being routed through a large number of hops. In the case where a second path or wavelength is not available, contention at an outbound link can be settled on a first-come, first-serve basis or on a priority basis. The information is presented to a regular IP router and then is reviewed by higher layer protocols, using retransmission when necessary.

Routing Example

An illustrative WDM circuit-switched backbone network 500 for communicating packets among end-users in certain large cities in the United States is shown in pictorial form in FIG. 5 -- network 500 is first discussed in terms of its conventional operation, that is, before the overlay of WDM optical tag switching in accordance with the present invention is presented.

With reference to FIG. 5, it is supposed that New York City is served by network element 501, Chicago is served by network element 502, ..., Los Angeles is served by network element 504, ..., and Minneapolis by network element 507. (Network elements may also be referred to a nodes in the sequel.) Moreover, NC&M 220 has logical connections (shown by dashed lines, such as channel 221 to network element 501 and channel 222 to network element 507) to all network elements 501-507 via physical layer optical supervisory channels; there is continuous communication among NC&M 220 and network elements 501-507. NC&M 220 periodically requests and receives

information about: (a) the general state of each network element (e.g., whether it is operational or shut down for an emergency); (b) the optical wavelengths provided by each network element (e.g., network element 501 is shown as being served by optical fiber medium 531 having wavelength W1 and optical fiber medium 532 having wavelength W2 which connect to network elements 502 (Chicago) and 505 (Boston), respectively); and (c) the ports which are served by the wavelengths (e.g., port 510 of element 501 is associated with an incoming client interface conveying packet 520, port 511 is associated with W1 and port 512 is associated with W2, whereas port 513 of element 502 is associated with W1).

Thus, NC&M 220 has stored at any instant the global information necessary to formulate routes to carry the incoming packet traffic by the network elements. Accordingly, periodically NC&M 220 determines the routing information in the form of, for example, global routing tables, and downloads the global routing tables to each of the elements using supervisory channels 221, 222, ... The global routing tables configure the ports of the network elements to create certain communication links. For example, NC&M 220 may determine, based upon traffic demand and statistics, that a fiber optic link from New York City to Los Angeles (network elements 501 and 504, respectively) is presently required, and the link will be composed, in series, of: W1 coupling port 511 of element 501 to port 513 in network element 502; W1 coupling port 514 of element 502 to port 515 of element 503; and W2 coupling port 516 of element 503 to port 517 of element 504. Then, input packet 520 incoming to network element 501 (New York City) and having a destination of network element 504 (Los Angeles) is immediately routed over this established link. At network element 504, the propagated packet is delivered as output packet 521 via client interface port 518.

In a similar manner, a dedicated path between elements 506 and 507 (St. Louis and Minneapolis, respectively) is shown as established using W2 between network elements 506 and 502, and W3 between elements 502 and 507.

Links generated in this manner -- as based upon the global routing tables -- are characterized by their rigidity, that is, it takes several seconds for NC&M 220 to determine the connections to establish the links, to download the connectivity information for the links, and establish the input and output ports for each network element. Each link has characteristics of a circuit-switched connection, that is, it is basically a permanent connection or a dedicated path or "pipe" for long intervals, and only NC&M 220 can tear down and re-establish a link in normal operation. The benefit of such a dedicated path is that traffic having an origin and a destination which maps into an already-established dedicated path can be immediately routed without the need for any set-up. On the other hand, the dedicated path can be, and most often is, inefficient in the sense that the dedicated path may be only used a small percentage of the time (e.g., 20%-50% over the set-up period). Moreover, switching device 430 (see FIG. 4) embedded in each network element which interconnects input and output ports has only a finite number of input/output ports. If the above scenario is changed so that link from St. Louis to Minneapolis is required and a port already assigned to the New York to Los Angeles link is to be used (e.g., port 514 of network element 502), then there is a time delay until NC&M 220 can respond and alter the global routing tables accordingly.

Now the example is expanded so that the subject matter in accordance with the principles of the present invention is overlaid on the above description. First, a parameter called the "tag-switched state" is introduced and its use in routing is discussed;

then, in the next paragraph, the manner of generating the tag-switch state is elucidated. The tag-switch state engenders optical tag switching.

NC&M 220 is further arranged so that it may assign the tag-switch state to each packet incoming to a network element from a client interface -- the tag-switch state is appended by Plug & Play module 132 and, for the purposes of the present discussion, the tag-switch state is commensurate with header 210 (see FIG. 2). The tag-switch state is computed by NC&M 220 and downloaded to each network element 501-507 in the form of a local routing table. With reference to FIG. 6, there is shown network element 501 and its embedded switch 601 in pictorial form. Also shown is incoming optical fiber 602, with delay loop 603, carrying packet 620 composed of header 210 and payload 211 – payload 211 in this case is packet 520 from FIG. 5. Fiber 6022 delivers a delayed version of packet 620 to network element 501. Also, a portion of the light energy appearing on fiber 602 is tapped via fiber 6021 and inputted to optical module 410 which processes the incoming packet 620 to detect header 210 -- header 210 for packet 620 is shown as being composed of the tag-switch state '11101011000', identified by reference numeral 615. Also shown in FIG. 6 is local look-up table 610, being composed of two columns, namely, "Tag-Switch State" (column 611), and "Local Address" (column 612). The particular tag-switch state for packet 620 is cross-referenced in look-up table 610 to determine the routing of the incoming packet. In this case, the tag-switch state for packet 620 is the entry in the fourth row of look-up table 610. The local switch address corresponding to this tag-switch state is "0111", which is interpreted as follows: the first two binary digits indicate the incoming port, and the second two binary digits indicate the output port. In this case, for the exemplary four-input, four-output switch, the incoming packet is to be routed from input port "01" to output port "11", so switch 601 is switched

accordingly (as shown). After the delay provided by fiber delay 603, the incoming packet on fiber 6022 is propagated onto fiber 604 via switch 601.

The foregoing description of tag-switch state indicates how it is used. The manner of generating the tag-switch state is now considered. NC&M 220, again on a periodic basis, compiles a set of local look-up tables for routing/switching the packet through each corresponding network element (such as table 610 for network element 501), and each look-up table is then downloaded to the corresponding network element. The generation of each look-up table takes into account NC&M 220's global knowledge of the network 500. For instance, if incoming packet 620 to network 501 is destined for network 504 (again, New York to Los Angeles), if port 510 is associated with incoming port "01" and serves fiber 602, and if outgoing port 511 is associated with outgoing port "11" and serves fiber 604, then NC&M 220 is able to generate the appropriate entry in look-up table 610 (namely, the fourth row) and download table 610 to network element 510. Now, when packet 520 is processed by electro-optical module 132 so as to add header 210 to packet 520 to create augmented packet 620, NC&M 220's knowledge of the downloaded local routing tables as well as the knowledge of the destination address embedded in packet 520 as obtained via module 132 enables NC&M 220 to instruct module 132 to add the appropriate tag-switch state as header 210 -- in this case '11101011000'.

It can be readily appreciated that processing a packet using the tag-switch state parameter is bursty in nature, that is, after switch 601 is set-up to handle the incoming tag-switch state, switch 601 may be returned to its state prior to processing the flow state. For example, switch 601 may have interconnected input port '01' to output port '10' prior to the arrival of packet 620, and it may be returned to the '0110' state after

processing (as determined, for example, by a packet trailer). Of course, it may be that the circuit-switched path is identical to the tag-switch state path, in which case there is no need to even modify the local route through switch 601 for processing the tag-switch state. However, if it is necessary to temporarily alter switch 601, the underlying circuit-switched traffic, if any, can be re-routed or re-sent.

As discussed so far, tag switching allows destination oriented routing of packets without a need for the network elements to examine the entire data packets. New signaling information -- the tag -- is added in the form of optical signal header 210 which is carried in-band within each wavelength in the multi-wavelength transport environment. This tag switching normally occurs on a packet-by-packet basis. Typically, however, a large number of packets will be sequentially transported towards the same destination. This is especially true for bursty data where a large block of data is segmented in many packets for transport. In such cases, it is inefficient for each particular network element to carefully examine each tag and decide on the routing path. Rather, it is more effective to set up a "virtual circuit" from the source to the destination. Header 210 of each packet will only inform continuation or ending of the virtual circuit, referred to as a flow state connection. Such an end-to-end flow state path is established, and the plug-and-play modules in the network elements will not disrupt such flow state connections until disconnection is needed. The disconnection will take place if such a sequence of packets have come to an end or another packet of much higher priority requests disruption of this flow state connection.

The priority aspect of the present invention is also shown with respect to FIG. 6. The local look-up table has a "priority level" (column 613) which sets forth the priority assigned to the tag-switching state. Also, header 210 has appended priority data

shown as the number '2' (reference numeral 616). Both the fourth and fifth row in the "tag-switch state" column 611 of table 610 have a local address of '0111.' If an earlier data packet used the entry in the fifth row to establish, for example, a virtual circuit or flow switching state, and the now another packet is processed as per the fourth row of column 611, the higher priority data ('2' versus '4', with '1' being the highest) has precedent, and the virtual circuit would be terminated.

Detailed Illustrative Embodiment

In order to achieve ultra-low latency IP over WDM tag switching, processing of the optical header at each optical switch must be kept to a minimum during the actual transmission of the optical packet. To achieve this end, a new signaling architecture and packet transmission protocol for performing optical WDM tag switching is introduced.

The signaling and packet transmission protocols decouple the slow and complex IP routing functions from the ultra-fast WDM switching functions. This decoupling is achieved via the setup up an end-to-end routing path which needs to be performed very infrequently. To send IP packets from a source to a destination, the following steps are executed:

(a) End-to-end routing path setup, where the IP layer software invokes the signaling protocol between the network elements and the NC&M to set up an end-to-end routing path for the IP packets. This step will also configure the WDM network elements along the routing path to support subsequent packet forwarding. The tags for optical tag switching to be inserted in the optical headers during actual packet transmission are also determined.

(b) Optical packet transmission, where the arrival of the optical packet triggers the local header processing which among other things looks up the output port for forwarding the packet on to the next hop based on the optical tag inside the optical header.

Although routing path setup involves invoking the routing function which is generally a slow and complicated procedure, it is performed prior to packet transmission handling, and hence it is not in the critical path that determines transmission latency.

Routing Path Setup

During routing path setup, the internal connection table of a WDM packet switch will be augmented with a tag-switch look-up table, and contains the pertinent packet forwarding information. In particular, in the interest of achieving ultra-low latency and hardware simplicity, the inventive scheme produces tag-switch states that remain constant along the flow path. For example, tag-switch assignments include the following techniques:

(1) Destination-based flow tag assignment -- In this scheme the destination, e.g. a suitable destination IP address prefix can be used as the tag-switch state in next hop look-up. In addition to having no need to modify the optical header, the same header can be used in the event of deflection routing.

(2) Route-based flow tag assignment -- In this scheme the tag-switch state assigned refers to the end-to-end route that is computed dynamically at the tag-switch state setup phase. The advantage of this scheme is that it can be specialized to meet the Quality-of-Service requirements for each individual tag-switched states.

Switching Conflict Resolution

The present-day lack of a viable optical buffer technology implies that conventional buffering techniques cannot be used to handle switching conflicts. As previously described, the invention embodiment utilizes fixed delay implemented by an optical fiber to allow switching to occur during this time delay, but not to achieve contention resolution as electrical buffers do in conventional IP routers. To resolve switching contentions, in accordance with the present invention, the following three methods are used:

(a) Limited wavelength interchange -- where a packet is routed through the same path but at a different wavelength. Since this wavelength conversion is utilized just to avoid the contention, it is not necessary that the network elements must possess the capability of converting to any of the entire wavelength channels. Rather, it is sufficient if they can convert some of the entire wavelength channels. This wavelength conversion converts both the signaling header and the data payload. Care must be taken to prevent a packet from undergoing too many wavelength conversions which will result in poor signal fidelity. A possible policy is to allow only one conversion, which and can easily be enforced by encoding the original wavelength in the optical header. This way an intermediate WDM switch will allow conversion if and only if it is carried on its original wavelength.

(b) Limited deflection routing -- where a packet may be deflected to a neighboring switching node from which it can be forwarded towards its destination. Care again must be taken to prevent a packet from being repeatedly deflected, thereby causing signal degradation, as well as wasting network bandwidth. A solution scheme is to

record a "timestamp" field in the optical header, and allow defections to proceed if and only if the recorded timestamp is no older than a maximum limit.

(c) Prioritized packet preemption -- where a newly arrived packet may preempt a currently transmitting packet if the arriving packet has a higher priority. The objective is to guarantee fairness to all packets so that eventually a retransmitted packet can be guaranteed delivery. In this scheme, each packet again has a timestamp field recorded in its optical header, and older packets have higher priority compared to newer packets. Furthermore a retransmitted packet assumes the timestamp of the original packet. This way, as a packet "ages," it increases in priority, and will eventually be able to preempt its way towards its destination if necessary.

It is noted that in all these schemes the optical header always remains constant as it moves around in the network. This is consistent with the desire to keep the optical switching hardware fast and simple. It is also possible to consider combinations of these schemes.

Routing Protocol

For a network the size of the NGI, centralized routing decisions are quite infeasible, so the approach needs to be generalized to distributed decision making. Hierarchical addressing and routing are used as in the case of IP routing. When a new connection is requested, NC&M 220 decides whether a WDM path is provisioned for this (source, destination) pair within the WDM-based network. If it is, the packets are immediately sent out on that (one-hop IP-level) path. If no such path is provisioned, NC&M 220 decides on an initial outbound link for the first WDM network element and a wavelength to carry the new traffic. This decision is based on the rest of the connections

in the network at the time the new connection was requested. NC&M 220 then uses signaling, through an appropriate protocol, to transfer the relevant information to the initial WDM network element to be placed in the signaling header. After the initial outbound link is determined, the rest of the routing decisions are taken at the individual NE's according to the optical signaling header information. This method ensures that the routing tables at each switching node and the signaling header processing requirements are kept relatively small. It also enables the network to scale easily in terms of switching nodes and network users. It is noted, too, that multiple WDM subnetworks can be interconnected together and each subnetwork will have its own NC&M.

When a path is decided upon, within a WDM NE, the optical switches can be set in that state (i) for the duration of each packet through the node and then revert back to the default state (called optical tag-switching), or (ii) for a finite, small amount of time (called flow switching). The former case performs routing on a regular packet-by-packet basis. The system resources are dedicated only when there is information to be sent and at the conclusion of the packet, these resources are available for assignment to another packet. The latter case is used for large volume bursty mode traffic. In this case, the WDM NE only has to read a flow state tag from the optical signaling header of subsequent packets arriving at the NE to be sure such packet is bound for the same destination, without the need to switch the switching device, and forward the payload through the already existing connection through the NE as previously established by the optical tag-switching.

The packets are self-routed through the network using the information in the signaling header of each packet. When a packet arrives at a switching node, the signaling header is read and either the packet is forwarded immediately through an

already existing flow state connection or a new appropriate outbound port is chosen according to the routing table. Routing tables in each node exist for each wavelength. If the packet cannot follow the selected outbound port because of contention with another packet (the selected outbound fiber is not free), the routing scheme will try to allocate a different wavelength for the same outbound port (and consequently the signal will undergo wavelength translation within the switching node). If no other eligible wavelength can be used for the chosen outbound port, a different outbound port may be chosen from another table, which lists secondary (in terms of preference) outbound links.

This routing protocol of the inventive technique is similar to the deflection routing scheme (recall the Background Section), where the session is deflected to some other outbound link (in terms of preference) if the preferred path cannot be followed. The packet is not allowed to be continuously deflected. In traditional routing protocols, a hop count is used to block a session after a specified number of hops. In the new scheme, in case no header regeneration is allowed at the switching nodes, then the hop count technique cannot be used. Alternatively, the optical signaling header characteristics (i.e., the signaling header's SNR) can be looked upon to decide whether a packet should be dropped.

IP Routing Algorithm in WDM layer

The technique used by NC&M 220 to determine the routing tables is based upon shortest path algorithms that route the packets from source to destination over the path of least cost. Specific cost criteria on each route, such as length, capacity utilization, hop count, or average packet delay can be used for different networks. The objective of the routing function is to have good performance (for example in terms of

low average delay through the network) while maintaining high throughput. Minimum cost spanning trees are generated having a different node as a root at each time, and the information obtained by these trees can then be used to set-up the routing tables at each switching node. If deflection routing as outlined above is implemented, the k-shortest path approach can be used to exploit the multiplicity of potential routing paths. This technique finds more than one shortest path, with the paths ranked in order of cost. This information can be inputted into the switching node routing tables, so that the outbound link corresponding to the minimum cost path is considered first, and the links corresponding to larger cost paths are inputted in secondary routing tables that are used to implement deflection routing.

Description of Plug-and-Play Modules

The present invention is based upon two types of Plug-and-Play modules to be attached to the WDM network elements. Introduction of these Plug-and-Play modules add optical tag switching capability to the existing circuit-switched network elements.

In FIG. 3, both header encoder 321 and header remover 322 were shown in high-level block diagram form; FIGS. 7 and 8 show, respectively, a more detailed schematic for both encoder 321 and remover 322.

In FIG. 7, IP packets or datagrams are processed in microprocessor 710 which generates each optical signaling header 210 for tag switching. Optical signaling header 210 and the original IP packet 211 are emitted from microprocessor 710 at baseband. Signaling header 210 is mixed in RF mixer 720 utilizing local oscillator 730. Both the mixed header from mixer 720 and the original packet 211 are combined in

combiner 740 and, in turn, the output of combiner 740 is encoded to an optical wavelength channel via optical modulator 760 having laser 750 as a source of modulation.

In FIG. 8, the optical channel dropping out of a network element is detected by photodetector 810 and is electrically amplified by amplifier 820. Normally, both photodetector 810 and the amplifier 820 have a frequency response covering only the data payload but not the optical signaling header RF carrier frequency provided by local oscillator 730. Low-pass-filter 830 further filters out any residual RF carriers. The output of filter 830 is essentially the original IP packet sent out by the originating IP router from the originating network element which has been transported through the network and is received by another IP router at another network element.

Block diagram 900 of FIG. 9 depicts the elements for the detection process effected by Plug-and-Play module 410 of FIG. 4 to convert optical signal 901, which carries both tag-switching signaling header 210 and the data payload 211, into baseband electrical signaling header 902. Initially, optical signal 901 is detected by photodetector 910; the output of photodetector 910 is amplified by amplifier 920 and filtered by high-pass filter 930 to retain only the high frequency components which carry optical signaling header 210. RF splitter 940 provides a signal to local oscillator 950, which includes feedback locking. The signal from local oscillator 950 and the signal from splitter 940 are mixed in mixer 960, that is, the high frequency carrier is subtracted from the output of filter 920 to leave only the information on tag-switching signaling header 210. In this process, local oscillator 950 with feedback locking is utilized to produce the local oscillation with the exact frequency, phase, and amplitude, so that the high frequency component is nulled during the mixing of this local oscillator signal and

the tag-switching signaling header with a high-frequency carrier. Low-pass filter 970, which is coupled to the output of mixer 960, delivers baseband signaling header 210 as electrical output signal 902.

The circuit diagram of FIG. 10 shows an example of a more detailed embodiment of FIG. 4. In FIG. 10, each header detector 1010, 1020, ..., 1030, ..., or 1040 processes information from each wavelength composing the optical inputs arriving on paths 1001, 1002, 1003, and 1004 as processed by demultiplexers 1005, 1006, 1007, and 1008, respectively; each demultiplexer is exemplified by the circuit 900 of FIG. 9. The processed information is grouped for each wavelength. Thus, for example, fast memory 1021 receives as inputs, for a given wavelength, the signals appearing on lead 1011 from header detector 1010, ..., and lead 1034 from header detector 1030. Each fast memory 1021-1024, such as a content-addressable memory, serves as an input to a corresponding tag switch controller 1031-1034. Each tag switch controller 1031-1034 also receives circuit-switched control signals from network element switch controller 420 of FIG. 4. Each tag switch controller intelligently chooses between the circuit switched control as provided by controller 420 and the tag switched information supplied by its corresponding fast memory to provide appropriate control signals the switching device 430 of FIG. 4.

Flow diagram 1100 of FIG. 11 is representative of the processing effected by each tag-switch controller 1031-1034. Using tag-switch controller 1031 as exemplary, inputs from circuit-switched controller 420 and inputs from fast memory 1021are monitored, as carried out by processing block 1110. If no inputs are received from fast memory 1021, then incoming packets are circuit-switched via circuit-switched controller 420. Decision block 1120 is used to determine if there are any inputs from fast memory

1021. If there are inputs, then processing block 1130 is invoked so that tag-switch controller 1031 can determine from the fast memory inputs the required state of switching device 430. Then processing block 1160 is invoked to transmit control signals from tag-switch controller 1031 to control switching device 430. If there are no fast memory inputs, then the decision block 1140 is invoked to determine if there are any inputs from circuit-switched controller 1140. If there are inputs from circuit-switched controller 420, then processing by block 1150 is carried out so that tag-switch controller 1031 determines from the inputs of circuit-switched controller 420 the required state of switching device 430. Processing block 1160 is again invoked by the results of processing block 1150. If there are no present inputs from circuit-switched controller 1140 or upon completion of procession block 1160, control is returned to processing block 1110.

By way of reiteration, optical tag-switching flexibly handles all types of traffic: high volume burst, low volume burst, and circuit switched traffic. This occurs by interworking of two-layer protocols of the tag-switched network control. Thus, the distributed switching control rapidly senses signaling headers and routes packets to appropriate destinations. When a long stream of packets reach the network element with the same destination, the distributed switching control establishes a flow switching connection and the entire stream of the packets are forwarded through the newly established connections.

A tag switching method scales graciously with the number of wavelengths and the number of nodes. This results from the fact that the distributed nodes process multi-wavelength signaling information in parallel and that these nodes incorporate

predicted switching delay in the form of fiber delay line. Moreover, the tag switching utilizes path deflection and wavelength conversion for contention resolution.

Optical Technology

Optical technologies span a number of important aspects realizing the present invention. These include optical header technology, optical multiplexing technology, optical switching technology, and wavelength conversion technology.

(a) Optical Header Technology

Optical header technology includes optical header encoding and optical header removal as discussed with respect to FIGS 3 and 4. In effect, optical header 210 serves as a signaling messenger to the network elements informing the network elements of the destination, the source, and the length of the packet. Header 210 is displaced in time compared to the actual data payload. This allows the data payload to have any data rates/protocols or formats.

As previously described with respect to FIGS. 7and 8, the header encoding is subcarrier based. This method allows header 210 to be separated in modulation frequency so that header detection can be relatively simple. Header 210, which precedes the data payload in the time domain, also has higher frequency carrier than the highest data rate. This allows reading of header 210, and eventually removal of header 210 without affecting the data payload.

(b) Optical Multiplexing Technology

Optical multiplexing may illustratively be implemented using the known silica arrayed waveguide grating structure. This waveguide grating structure has a

number of unique advantages including: low cost, scalability, low loss, uniformity, and compactness.

(c) Optical Switching Technology

Fast optical switches are essential to achieving packet routing without requiring excessively long fiber delay as a buffer.

Micromachined Electro Mechanical Switches offer the best combination of the desirable characteristics: scalability, low loss, polarization insensitivity, fast switching, and robust operation. Recently reported result on the MEM based Optical Add-Drop Switch achieved 9 microsecond switching time

(d) Wavelength Conversion Technology

Wavelength conversion is resolves packet contention without requiring path deflection or packet buffering. Both path deflection and packet buffering cast the danger of skewing the sequences of a series of packets. In addition, the packet buffering is limited in duration as well as in capacity, and often requires non-transparent methods. Wavelength conversion, on the other hand, resolves the blocking by transmitting at an alternate wavelength through the same path, resulting in the identical delay. Illustratively, a WSXC with a limited wavelength conversion capability is deployed.

Although various embodiments which incorporate the teachings of the present invention have been shown and described in detail herein, those skilled in the art can readily devise many other varied embodiments that still incorporate these teachings.

CLAIMS

What is claimed is:

1. A method for propagating a data payload from an input network element to an output network element in a wavelength division multiplexing (WDM) network composed of a plurality of network elements, the data payload having a given format and protocol, the method comprising the steps of

generating and storing a local routing look-up table in each of the network elements, each local routing table determining a local route through the associated one of the network elements,

adding an optical header to the data payload prior to inputting the data payload to the input network element, the header having a format and protocol and being indicative of the local route through each of the network elements for the data payload and the header, the format and protocol of the data payload being independent of the format and protocol of the header,

optically determining the header at the network elements as the data payload and header propagate through the WDM network,

selecting the local route for the data payload and the header through the network elements as determined by looking up the header in the corresponding local routing table, and

routing the data payload and the header through the network elements in correspondence to the selected route.

2. The method as recited in claim 1 wherein the optical header includes a tag-switch state for routing the optical header and the data payload through the network elements, and the step of adding an optical header includes the steps of determining and inserting in the optical header an appropriate tag-switch state for routing the optical header and the data payload from the input network element to the output network element through the network elements.

3. The method as recited in claim 2 wherein the optical header further includes priority data for use in resolving route contentions as the optical header and the data payload propagate through the network elements, the step of determining and storing a local routing table includes the step of associating a priority level with each tag-switch state, the step of adding the optical header includes the step of inserting in the optical header appropriate priority data for the data payload, and the step of selecting includes the step of determining the local route based upon the priority data and the priority level.

4. The method as recited in claim 1 wherein the step of adding the optical header to the data payload includes the step of placing the optical header ahead of the data payload in time.

5. The method as recited in claim 1 wherein the optical header and the data payload are initially generated at baseband, and the step of adding the optical header to the data payload further includes the steps of

frequency shifting the baseband optical header to a frequency band above the frequency band of the baseband data payload,

combining the frequency-shifted baseband optical header and the baseband data payload to form a composite frequency signal, and

optically modulating the composite frequency signal using an optical source of a given wavelength to produce an optical signal to propagate the header and the data payload through the WDM network.

6. The method as recited in claim 5 wherein the step of optically determining the header at each of the network elements includes the steps of

photo-detecting the optical header to produce a detected signal,

locking onto the detected signal with a local locking oscillator to produce a locked signal, and

mixing the detected signal and the locked signal to produce a baseband signal representative of the header at baseband.

7. The method as recited in claim 1 wherein the step of routing includes the step of resolving contentions for the selected route.

8. The method as recited in claim 7 wherein the step of resolving contentions includes the step of routing over an alternate route determined with reference to the selected route.

9. The method as recited in claim 7 wherein the step of resolving contentions includes the step of routing over an alternate wavelength determined with respect to a wavelength used for the selected route.

10. A method for propagating a sequence of data payloads from an input network element to an output network element in a wavelength division multiplexing (WDM) network composed of a plurality of network elements, each of the data payloads having a given format and protocol, the method comprising the steps of

generating and storing a local routing look-up table in each of the network elements, each local routing table determining a local route through the associated one of the network elements,

adding an optical header to each of the data payloads prior to inputting the data payloads to the input network element, the header having a format and protocol and being indicative of the local route through each of the network elements for each of the data payloads and its corresponding header, the format and protocol of each of the data payloads being independent of the format and protocol of its corresponding header,

optically determining the header at the network elements as each of the data payloads and its corresponding header propagate through the WDM network,

selecting the local route for the first of the data payloads and its corresponding header through the network elements as determined by looking up the header in the corresponding local routing table,

routing the first of the data payloads and its corresponding header through the network elements in correspondence to the selected route, and

routing subsequent ones of the data payloads in the sequence through the local route selected for the first of the data payloads.

11. The method as recited in claim 10 wherein each step of routing includes the step of resolving contentions for the selected route.

12. The method as recited in claim 11 wherein the step of resolving contentions includes the step of routing over an alternate route determined with reference to the selected route.

13. The method as recited in claim 11 wherein the step of resolving contentions includes the step of routing over an alternate wavelength determined with respect to a wavelength used for the selected route.

14. A method for propagating a data payload arriving at an input network element onto a wavelength division multiplexing (WDM) network composed of a plurality of network elements, the data payload having a given format and protocol, the method comprising the steps of

generating an optical header associated with the data payload, the header having a format and protocol and being indicative of a local route through each of the network elements for the data payload and the header, the format and protocol of the data payload being independent of the format and protocol of the header, and

adding the optical header to the data payload prior to inputting the data payload to the input network element.

15. The method as recited in claim 14 wherein the optical header and the data payload are initially generated at baseband, and the step of adding the optical header to the data payload further includes the steps of

frequency shifting the baseband optical header to a frequency band above the frequency band of the baseband data payload,

combining the frequency-shifted baseband optical header and the baseband data payload to form a composite frequency signal, and

optically modulating the composite frequency signal using an optical source of a given wavelength to produce an optical signal for propagating the header and the data payload.

16. A method for transferring a header and a data payload from the input to the output of each particular network element in a wavelength division multiplexing (WDM) network composed of a plurality of network elements, the data payload having a given format and protocol independent of those of the header, the method comprising the steps of

generating and storing a local routing look-up table in the particular network element, the local routing table determining a local route through the particular network element,

optically determining the header as the data payload and header arrive at the input to the particular network element,

selecting the local route for the data payload and the header through the particular network element as determined by looking up the header in the local routing table, and

routing the data payload and the header through the particular network element in correspondence to the selected route.

17. The method as recited in claim 16 wherein the step of optically determining the header at each of the network elements includes the steps of

photo-detecting the optical header to produce a detected signal,

locking onto the detected signal with a local locking oscillator to produce a locked signal, and

mixing the detected signal and the locked signal to produce a baseband signal representative of the header at baseband.

18. A system, in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source device in the electrical layer and destined for a destination device in the electrical layer, the data payload having a given format and protocol, the system comprising

a first type of optical header module, coupling the source device and the WDM network, for adding an optical header ahead of the data payload prior to inputting

the data payload to the WDM network, the header being indicative of a local route through the network elements for the data payload and the header, the format and protocol of the data payload being independent of those of the header, and

a second type of optical header module, appended to each of the network elements, including means for storing a local routing look-up table in a corresponding one of the network elements, each local routing table determining a routing path through the corresponding one of the network elements, means for optically determining the header at the corresponding one of the network elements as the data payload and header propagate over the WDM network, means for selecting the local route for the data payload and the header through the corresponding one of the network elements as determined by looking up the header in the corresponding local routing table, and means for routing the data payload and the header through the corresponding one of the network elements in correspondence to the selected route.

19. The system as recited in claim 18 wherein another of the first type of optical header module couples the WDM network to the destination device, and the first type of optical header module further includes means for removing the header from the data payload before delivery to the destination device.

20. An optical header module, in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source device in the electrical layer and destined for a destination device in the electrical

layer, the data payload having a given format and protocol, the optical header module, coupling the source device and the WDM network, including means for generating an optical header associated with the data payload, the header having a format and protocol and being indicative of a local route through each of the network elements for the data payload and the header, the format and protocol of the data payload being independent of the format and protocol of the header, and means for adding the optical header to the data payload prior to inputting the data payload to the input network element.

21. The system as recited in claim 20 wherein the optical header includes a tag-switch state for routing the optical header and the data payload through the network elements, and the means for adding an optical header includes the means for determining and for inserting in the optical header an appropriate tag-switch state to route the optical header and the data payload through the network elements.

22. The system as recited in claim 20 wherein the means for adding the optical header to the data payload includes means for placing the optical header ahead of the data payload in time.

23. The system as recited in claim 22 wherein the optical header and the data payload are initially generated at baseband, and the means for adding the optical header to the data payload further includes

means for frequency shifting the baseband optical header to a frequency band above the frequency band of the baseband data payload,

means for combining the frequency-shifted baseband optical header and the baseband data payload to form a composite frequency signal, and

means for optically modulating the composite frequency signal to produce an optical signal for propagating the header and the data payload at a given wavelength.

24. An optical header processor, in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source device in the electrical layer and being destined for a destination device in the electrical layer, the data payload having a given format and protocol, the optical header processor module, associated with each of the network elements, comprising

means for storing a local routing look-up table in each corresponding one of the network elements, each local routing table determining a routing path through the corresponding one of the network elements,

means for optically determining the header at the corresponding one of the network elements as the data payload and header propagate over the WDM network,

means for selecting the local route for the data payload and the header through the corresponding one of the network elements as determined by looking up the header in the corresponding local routing table, and

means for routing the data payload and the header through the corresponding one of the network elements in correspondence to the selected route.

25. The header processor as recited in claim 24 wherein the means for optically determining the header at each of the network elements includes

means for photo-detecting the optical header to produce a detected signal,

a local locking oscillator for locking onto the detected signal to produce a locked signal, and

means for mixing the detected signal and the locked signal to produce a baseband signal representative of the header at baseband.

26. A system, in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source device in the electrical layter and being destined for a destination device in the electrical layer, the data payload having a given format and protocol, the network further including a network manager coupled to the network elements for determining circuit-switched routes through the network, with each of the network elements including (i) a switching device, and (ii) a circuit-switched controller, responsive to the network manager, for controlling the switching device based upon inputs from the network manager to established circuit-switched routing paths through the WDM network, the system comprising

a first type of optical header module, coupling the source device and the WDM network, for adding an optical header ahead of the data payload prior to inputting the data payload to the WDM network, the header being indicative of a local route through the network elements for the data payload and the header, the format and protocol of the data payload being independent of those of the header, and

a second type of optical header module, responsive to the network manager and the circuit-switched controller and coupled to the switching device, including means for storing a local routing table in each network element as provided by the network manager, each local routing table determining a routing path through each network element, means for optically determining the header at each network element as the data payload and header propagate over the WDM network, means for selecting the local route for the data payload and the header through each network element as determined by looking up the header in the corresponding local routing table, and means for routing the data payload and the header through each network element in correspondence to the selected route by processing inputs from the circuit-switched controller and the local routing table to control the switching device.

27. The system as recited in claim 26 further including means, interposed before the switching device, for delaying the delivery of the data payload and the header to the switching device for a pre-determined interval.

28. The system as recited in claim 26 wherein

the means for optically determining the header at each network element further includes a demultiplexer for demodulating the header propagating over the WDM network to a baseband header,

the means for selecting includes a fast memory, responsive to the demultiplexer, for determining route information contained in the baseband header, and

the means for routing includes a tag-switch controller, coupled to the local routing table and responsive to the fast memory and the circuit-switched controller, for controlling the switching device.

29. A system, in combination with (a) an electrical layer; and (b) an optical layer composed of a wavelength division multiplexing (WDM) network including a plurality of network elements, for propagating a data payload generated by a source device in the electrical layer and destined for a destination device in the electrical layer, the data payload having a given format and protocol, the system comprising

a first type of optical header module, coupling the source device and the WDM network, for adding an optical header ahead of the data payload prior to inputting the data payload to the WDM network, the header being indicative of a local route through the network elements for the data payload and the header, the format and protocol of the data payload being independent of that of the header, and

a second type of optical header module, appended to each of the network elements, including means for storing a local routing table in a corresponding one of the network elements, each local routing table determining a routing path through the corresponding one of the network elements, means for optically determining the header at the corresponding one of the network elements as the data payload and header propagate over the WDM network, means for selecting the local route for the data payload and the header through the corresponding one of the network elements as determined by looking up the header in the corresponding local routing table, means for routing the data payload and the header through the corresponding one of the network elements in correspondence

to the selected route, and means for maintaining the selected route for each subsequent consecutive header having the same local route.

FIG. 1

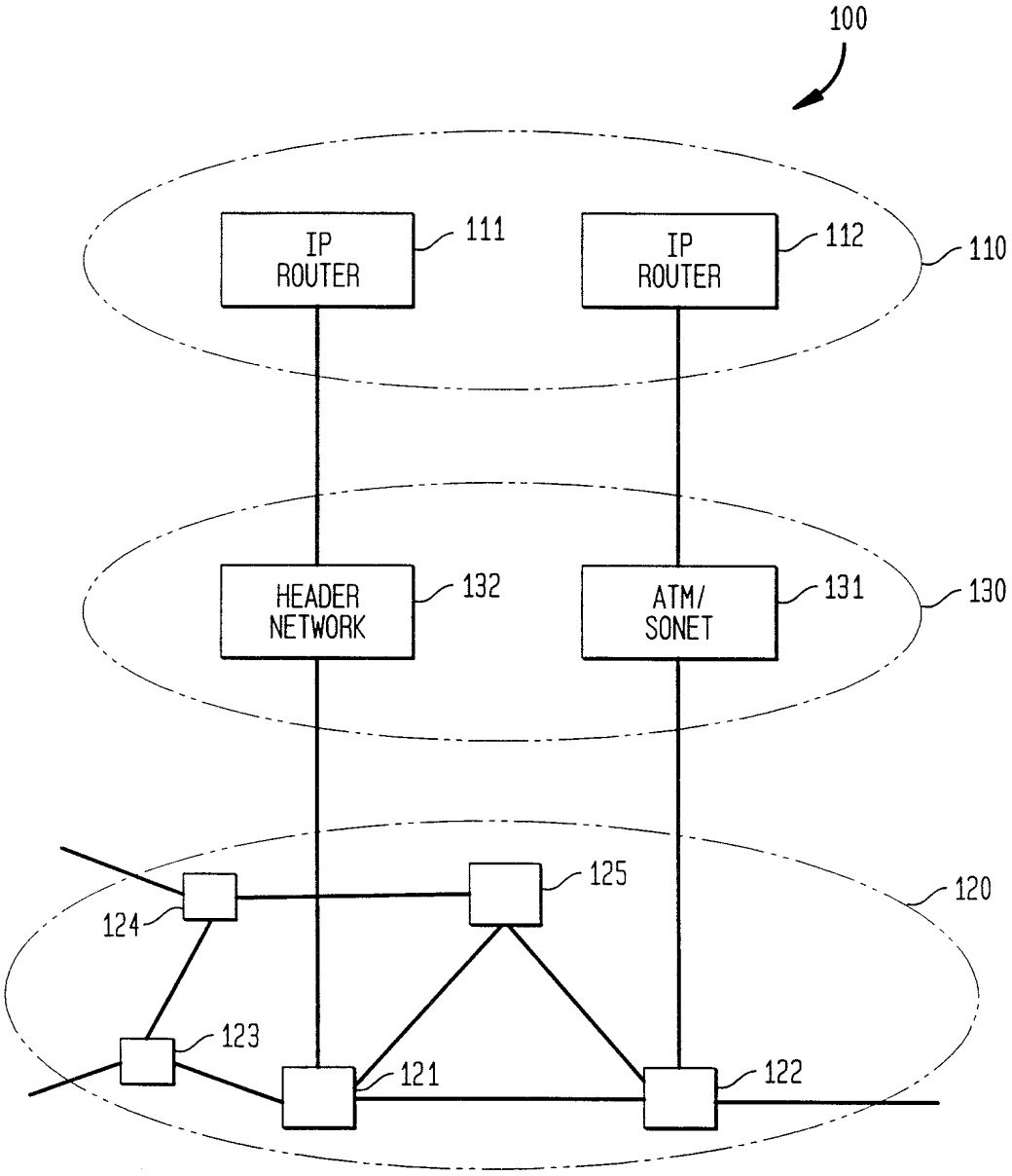

100
IP ROUTER
111
IP ROUTER
112
110
HEADER NETWORK
132
ATM/ SONET
131
130
124
125
120
123
121
122

FIG. 2

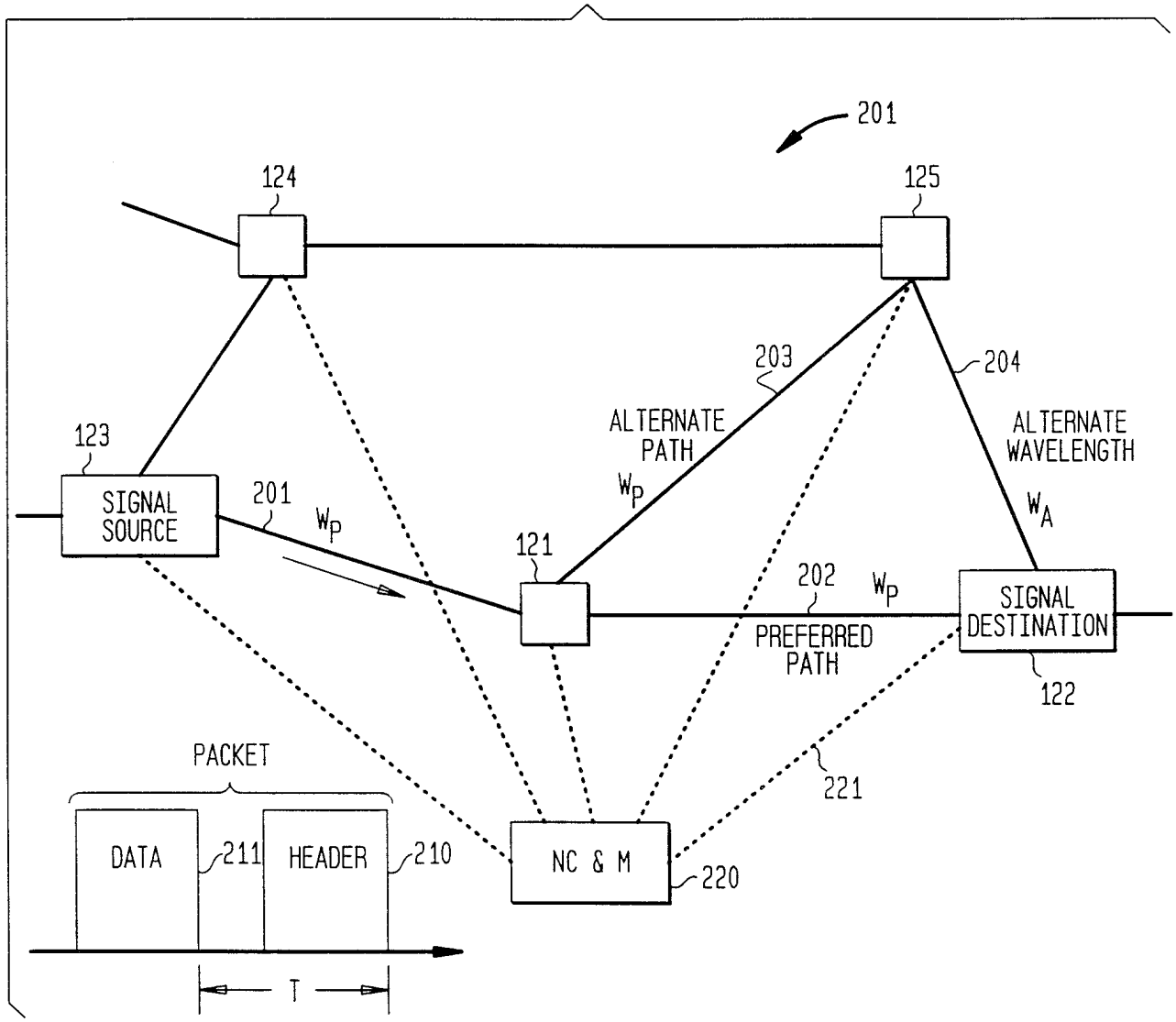

201
124
125
203
204
ALTERNATE PATH
ALTERNATE WAVELENGTH
123
SIGNAL SOURCE
201
W_P
W_P
W_A
121
202
W_P
SIGNAL DESTINATION
PREFERRED PATH
122
PACKET
DATA
211
HEADER
210
NC & M
220
221
T

FIG. 3

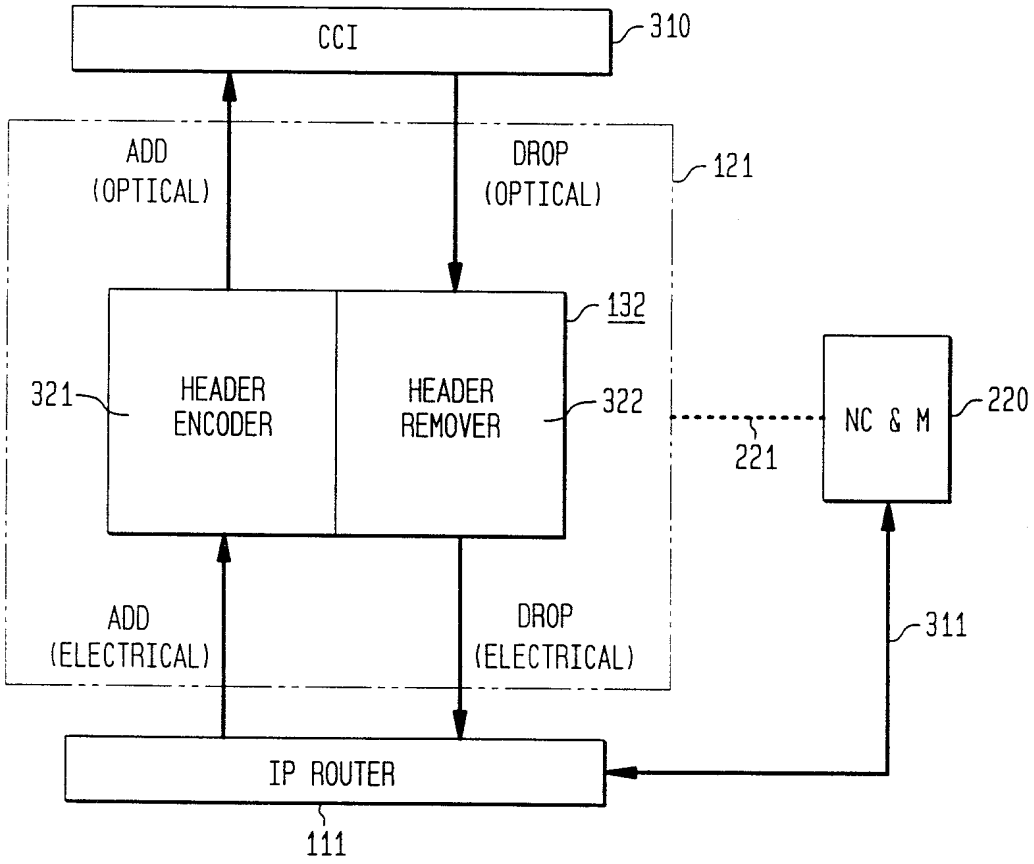

CCI
310
ADD
(OPTICAL)
DROP
(OPTICAL)
121
132
321
HEADER
ENCODER
HEADER
REMOVER
322
NC & M
220
221
ADD
(ELECTRICAL)
DROP
(ELECTRICAL)
311
IP ROUTER
111

FIG. 4

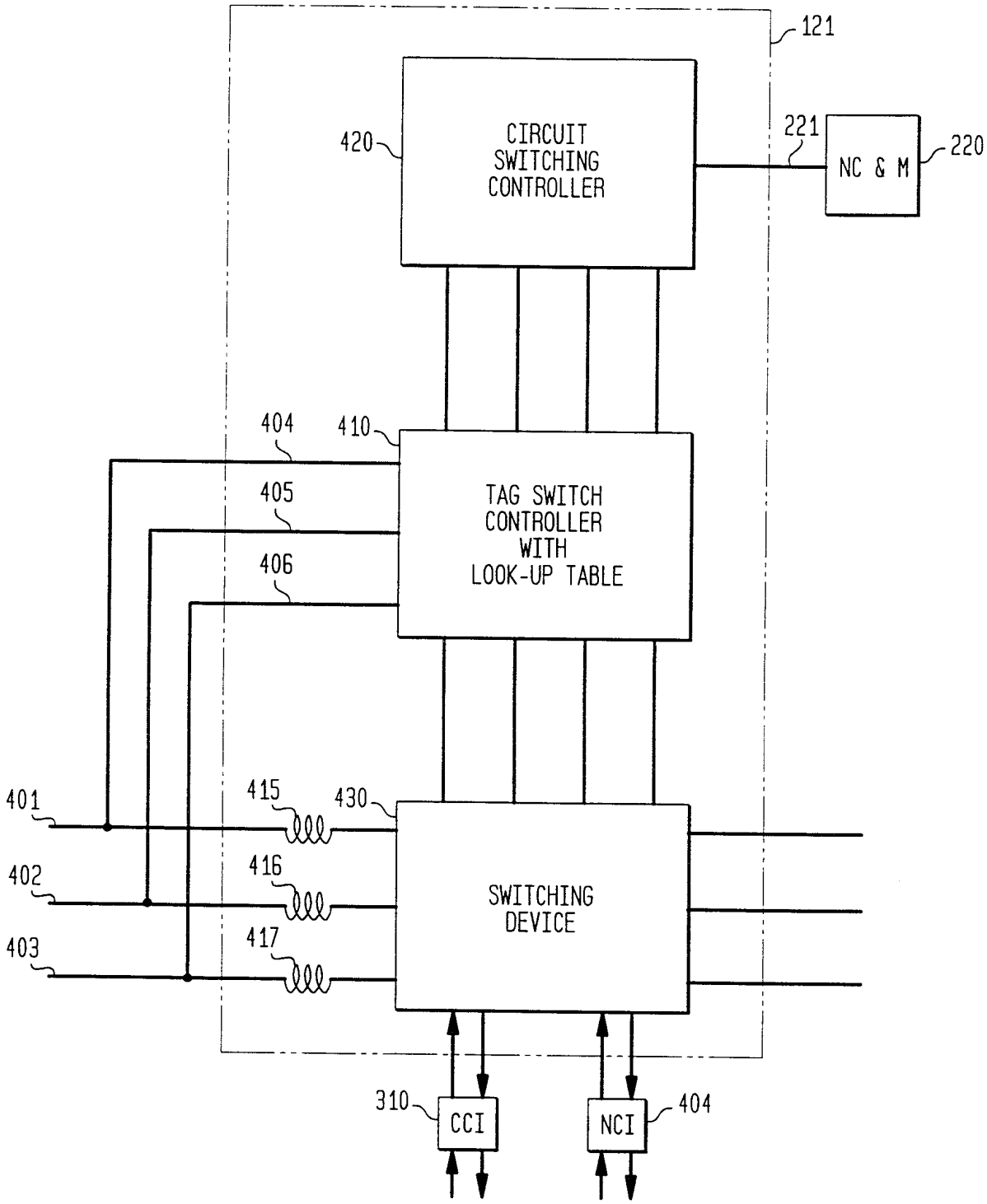

121
420
CIRCUIT SWITCHING CONTROLLER
221
NC & M
220
404
410
405
406
TAG SWITCH CONTROLLER WITH LOOK-UP TABLE
401
402
403
415
416
417
430
SWITCHING DEVICE
310
CCI
NCI
404

FIG. 5

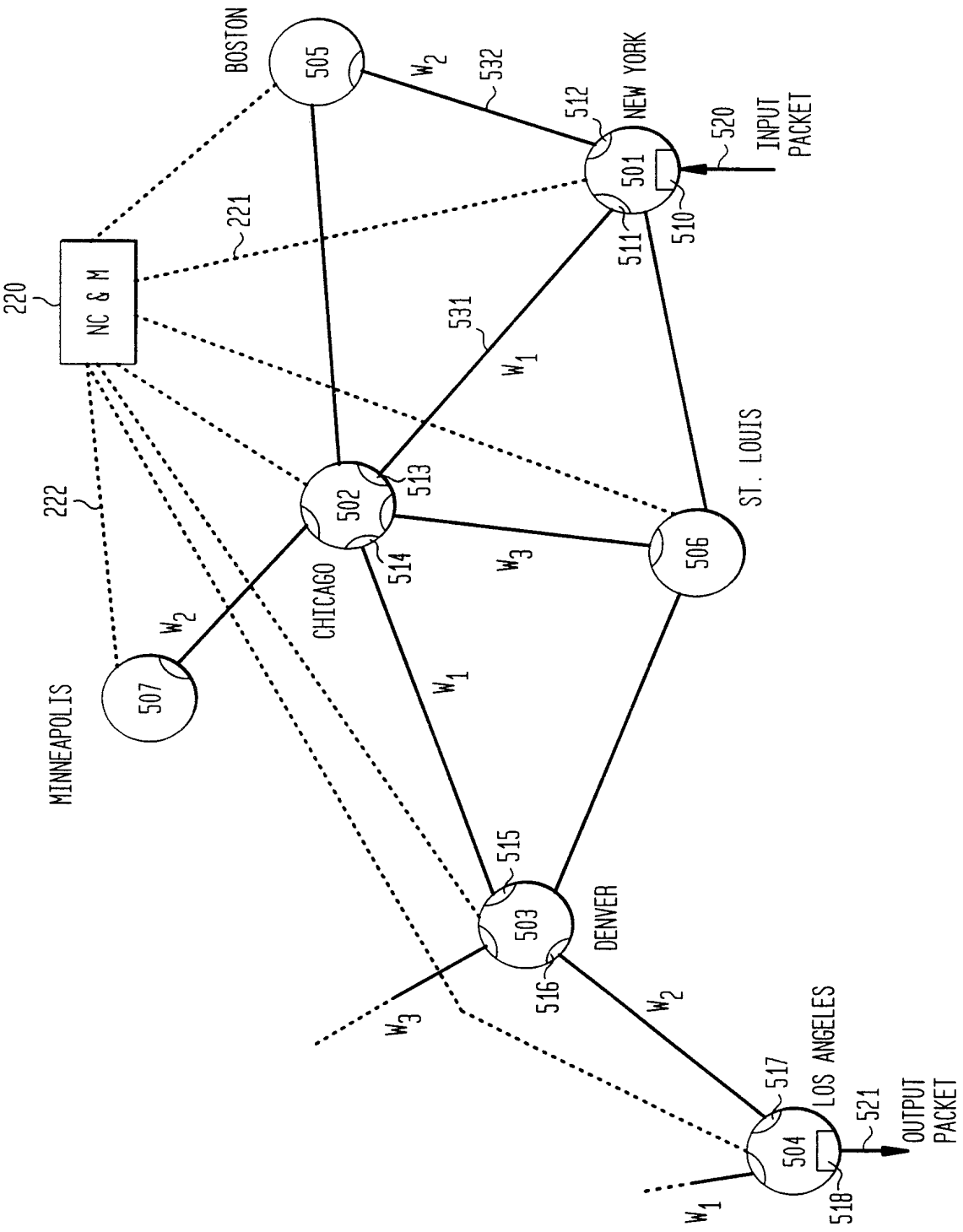

BOSTON
505
W_2
532
512
NEW YORK
520
INPUT
PACKET
501
510
511
221
220
NC & M
531
W_1
222
513
502
ST. LOUIS
506
514
W_3
CHICAGO
MINNEAPOLIS
507
W_2
W_1
515
503
DENVER
516
W_3
W_2
517
LOS ANGELES
521
OUTPUT
PACKET
504
518
W_1

FIG. 6

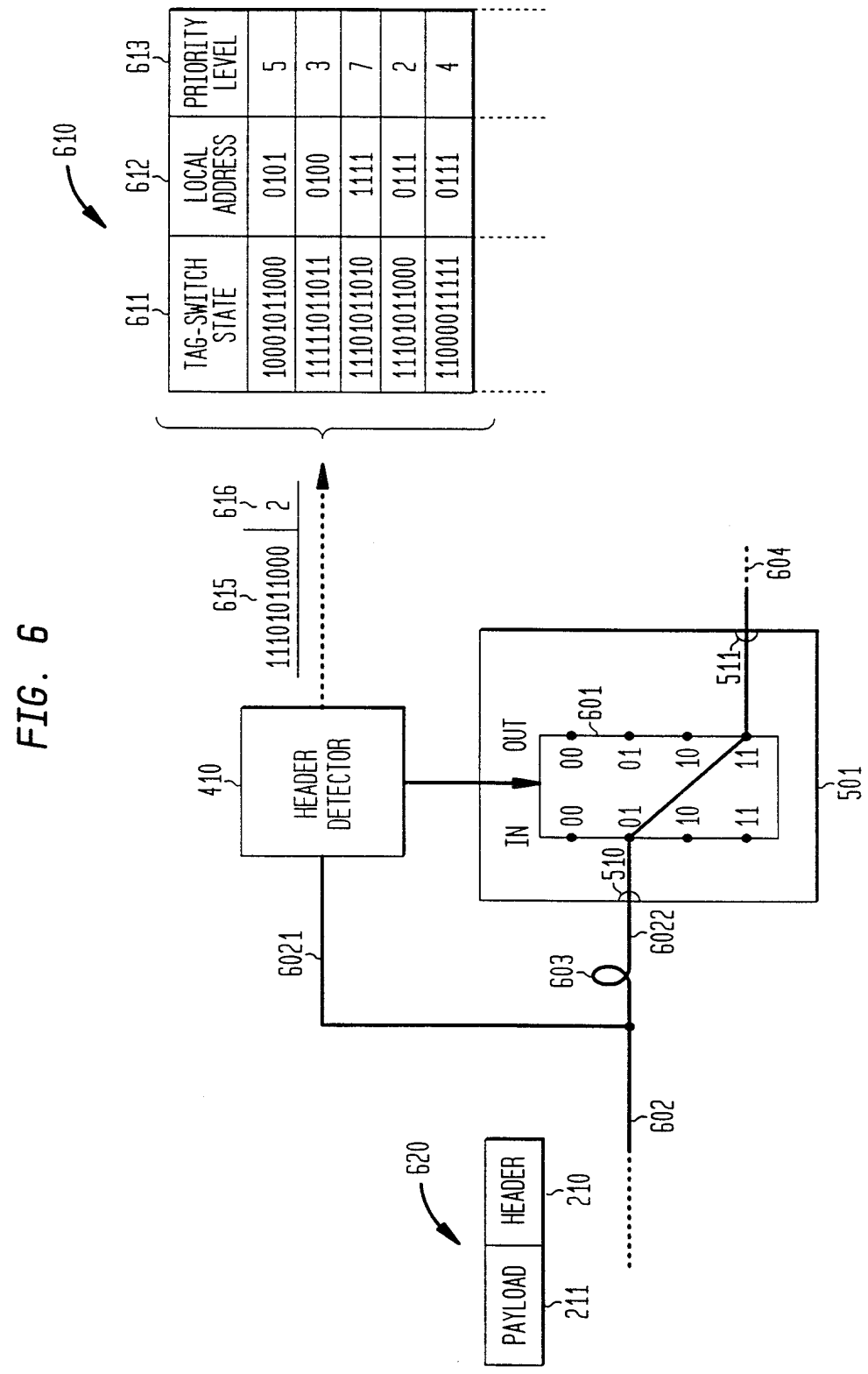

610
611
612
613
TAG-SWITCH STATE
LOCAL ADDRESS
PRIORITY LEVEL
10001011000
0101
5
11111011011
0100
3
11101011010
1111
7
11101011000
0111
2
11000011111
0111
4
615
11101011000
616
2
410
HEADER DETECTOR
IN
OUT
601
00
01
10
11
501
511
604
510
6021
6022
603
602
620
HEADER
210
PAYLOAD
211

FIG. 7

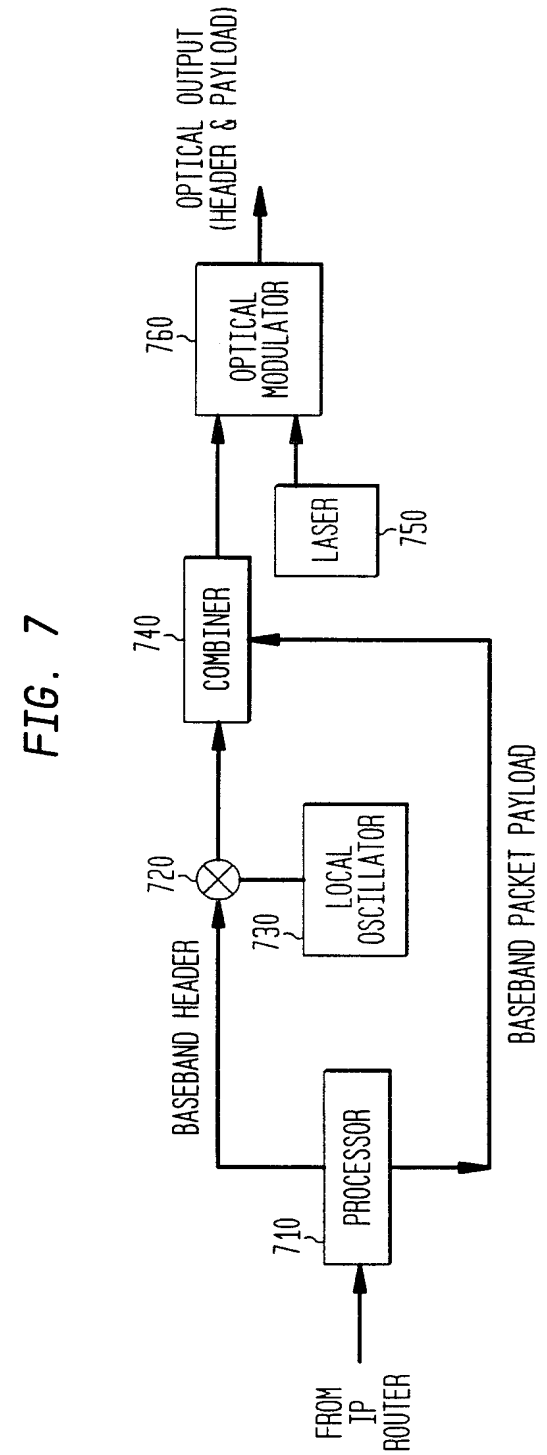

OPTICAL OUTPUT
(HEADER & PAYLOAD)
760
OPTICAL MODULATOR
LASER
750
740
COMBINER
720
LOCAL OSCILLATOR
730
BASEBAND HEADER
BASEBAND PACKET PAYLOAD
710
PROCESSOR
FROM IP ROUTER

FIG. 8

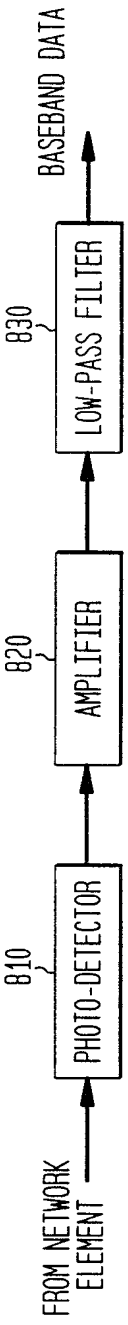
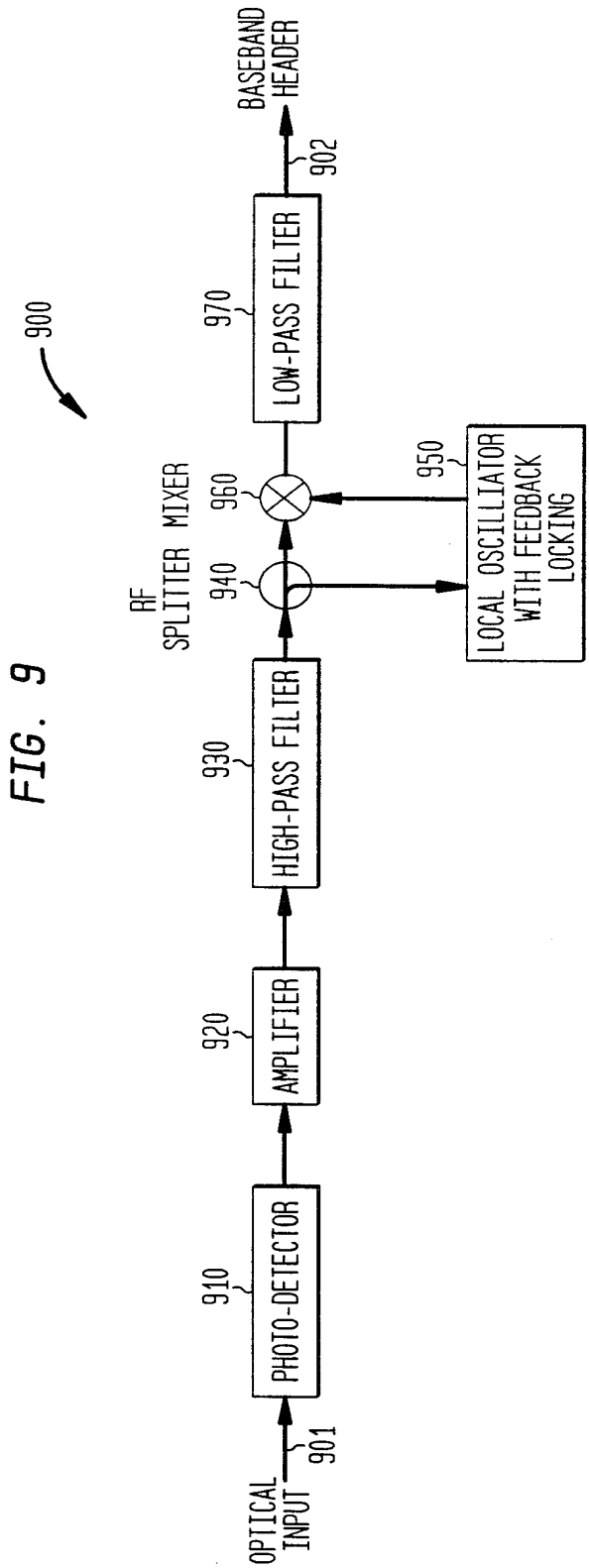

BASEBAND DATA
830
LOW-PASS FILTER
820
AMPLIFIER
810
PHOTO-DETECTOR
FROM NETWORK ELEMENT

FIG. 9


900
OPTICAL INPUT
901
910
PHOTO-DETECTOR
920
AMPLIFIER
930
HIGH-PASS FILTER
RF SPLITTER
940
MIXER
960
950
LOCAL OSCILLIATOR WITH FEEDBACK LOCKING
970
LOW-PASS FILTER
902
BASEBAND HEADER

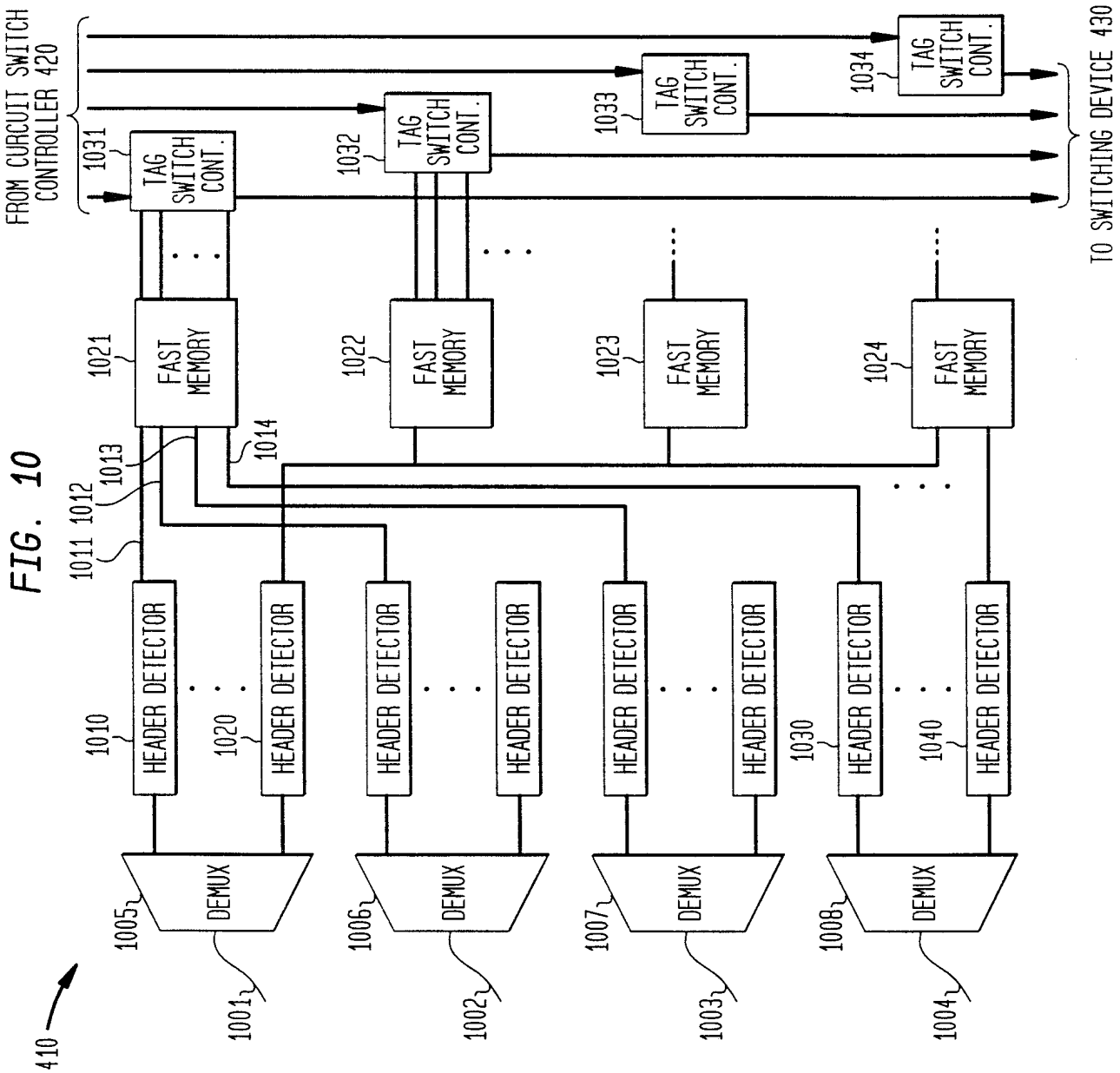
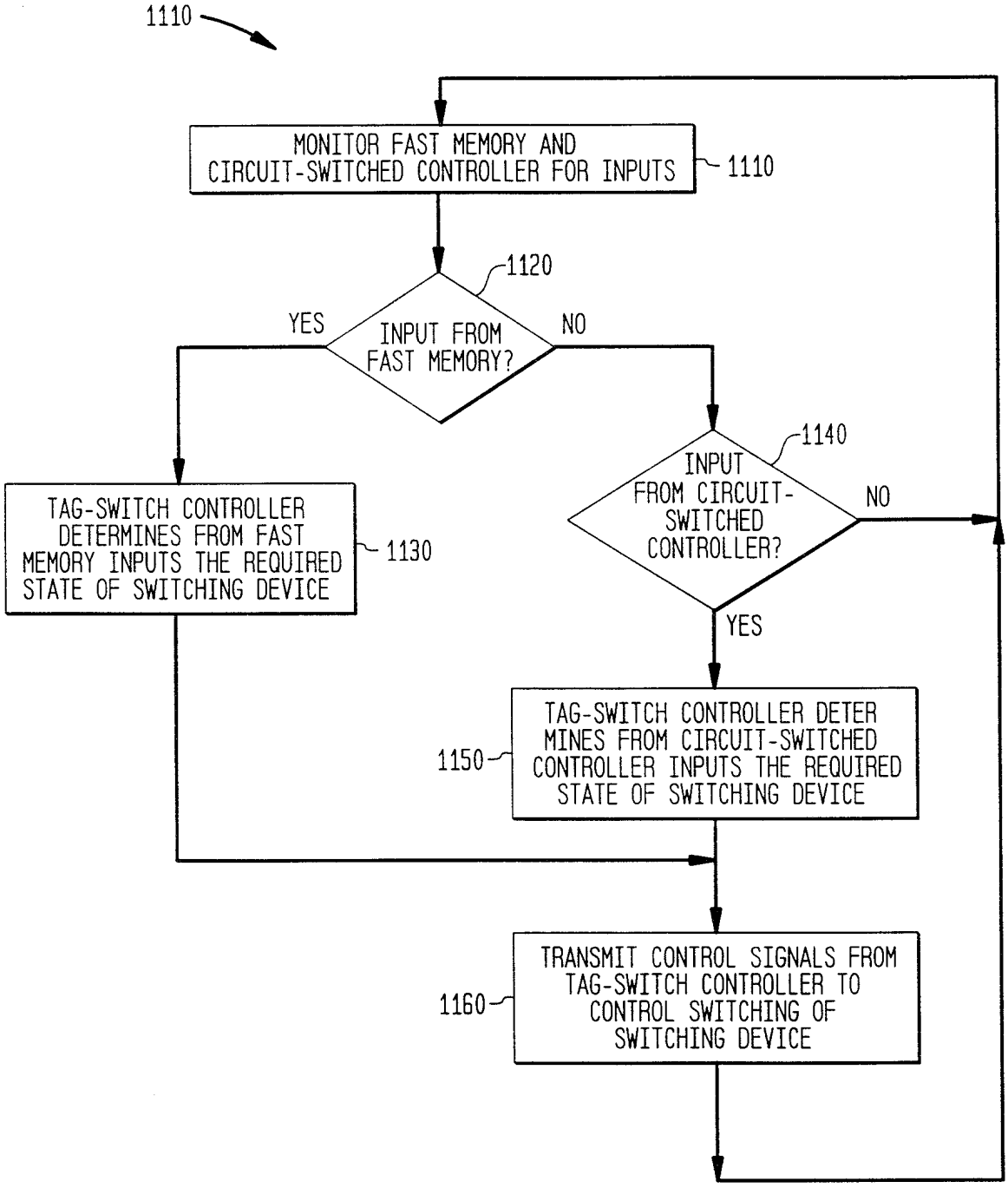

FROM CURCUIT SWITCH CONTROLLER 420
TO SWITCHING DEVICE 430
TAG SWITCH CONT.
1031
1032
1033
1034
FAST MEMORY
1021
1022
1023
1024
1011
1012
1013
1014
HEADER DETECTOR
1010
1020
1030
1040
DEMUX
1005
1006
1007
1008
1001
1002
1003
1004
410

FIG. 10

FIG. 11


1110
MONITOR FAST MEMORY AND CIRCUIT-SWITCHED CONTROLLER FOR INPUTS
1110
1120
INPUT FROM FAST MEMORY?
YES
NO
TAG-SWITCH CONTROLLER DETERMINES FROM FAST MEMORY INPUTS THE REQUIRED STATE OF SWITCHING DEVICE
1130
1140
INPUT FROM CIRCUIT-SWITCHED CONTROLLER?
NO
YES
1150
TAG-SWITCH CONTROLLER DETER MINES FROM CIRCUIT-SWITCHED CONTROLLER INPUTS THE REQUIRED STATE OF SWITCHING DEVICE
1160
TRANSMIT CONTROL SIGNALS FROM TAG-SWITCH CONTROLLER TO CONTROL SWITCHING OF SWITCHING DEVICE

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/14979

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :H04J 14/02

US CL : 359/123-124,128; 370/392-393,471,474

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 359/123-124,128; 370/392-393,471,474

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,488,501 A (BARNSLEY) 30 January 1996. See Fig.1	1-2,4,7-14, 16,20-22, 24

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 17 SEPTEMBER 1999	Date of mailing of the international search report 19 OCT 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer For KINFE-MICHAEL NEGASH Telephone No. (703) 305-4932

Form PCT/ISA/210 (second sheet)(July 1992) ★